Analyse de documents pédagogiques en vue de leur annotation

Boutheina Smine*, Rim Faiz**, Jean-Pierre Desclés***

* LARODEC, ISG de Tunis, 2000 Le Bardo, Tunisie
boutheina.smine@yahoo.fr
**LARODEC, IHEC de Carthage, 2016 Carthage Présidence, Tunisie
Rim.Faiz@ihec.rnu.tn
***LaLIC, Université Paris-Sorbonne, 28, rue Serpente, 75006 Paris, France
Jean-pierre.Descles@paris4.sorbonne.fr

Résumé. L'utilisation des documents pédagogiques, disponibles sur le web, devient de plus en plus large tant pour l'enseignant qui a besoin de préparer son support de cours que pour l'étudiant qui désire, par exemple, s'autoformer. La description d'un document pédagogique, en l'alimentant par des métadonnées, s'avère une solution qui confère une valeur ajoutée au document afin d'expliciter des informations placées dans ce document. Dans cette optique, nous proposons une méthode d'annotation de documents pédagogiques selon différents points de vue, qui est basée sur l'analyse sémantique des éléments discursifs du texte.

1 Introduction

Le besoin des utilisateurs, en documents pédagogiques, est sans cesse croissant en vue de leur réutilisation dans un processus d'apprentissage, de formation ou d'enseignement. De tels documents peuvent être disponibles sur le web, ce qui soulève un certain nombre de questions : comment retrouver ces documents ? Est ce qu'ils répondent au besoin spécifique de chaque utilisateur ? D'où l'utilité de décrire des documents pédagogiques en les alimentant par des métadonnées afin de faciliter l'exécution de tâches ultérieures sur ces documents. Toutefois, un document pédagogique possède des caractéristiques qui le distinguent des autres types de documents (*article de Presse, document publicitaire, etc.*) comme, par exemple : L'hétérogénéité des données contenues dans le document (*Exercices, Exemples, Tableaux, Figures, etc.*), la structure du document (elle peut varier d'un document à un autre). En d'autres termes, les documents pédagogiques peuvent être sous forme d'un support de cours, de travaux dirigés, de présentation ou encore sous forme de site web contenant des pages html. Ces caractéristiques doivent être prises en compte pour garantir une meilleure analyse des documents pédagogiques et ainsi une meilleure annotation de ces documents. Proposer une méthode d'annotation sémantique de documents pédagogiques s'avère d'une utilité considérable, cela facilitera le repérage des informations dont a besoin l'utilisateur à partir de documents pédagogiques. Ceci fait l'objet de notre travail.

Dans la suite, nous positionnons notre contribution par rapport aux travaux existants, puis nous présentons la notion de point de vue. Nous décrivons, ensuite, notre méthode d'annotation de documents pédagogiques qui est validée par le système ADoP présenté dans la cinquième section.

2 Présentation de quelques systèmes d'annotation

Il existe plusieurs systèmes d'annotation de documents pédagogiques utilisés dans le cadre du web sémantique. Parmi eux, nous distinguons ceux qui proposent une annotation semi-automatique du contenu du document pédagogique en utilisant le schéma RDF. Citons comme exemple, le système QBLS, proposé par Dehors et al. (2005), qui permet d'annoter le cours en se référant à une ontologie pédagogique composée de fiches (*définition, exemple, énoncé, procédure, solution, etc.*) et à des ressources pédagogiques abstraites (*cours, thème, notion, question*). Cependant, le système nécessite un enrichissement de son ontologie. Nous retrouvons la même technique d'annotation dans KATIA (Bodain**,** 2006) où l'annotation de chaque élément du texte est réalisée en lui correspondant une classe de l'ontologie sélectionnée directement à partir du web. Pour une mise à jour ultérieure, les annotations sont inscrites dans un fichier externe RDF relié à la page HTML concernée par ces annotations. Il existe d'autres systèmes qui sont caractérisés par leur annotation des cours en se référant à une ontologie d'un domaine particulier. Par exemple, Buffa et al. (2005), à travers la plateforme TRIAL SOLUTION, proposent des annotations de trois types : (1) les « types » (*définition, théorème, explication, etc.*), (2) les « mots-clés » (3) les « relations » avec d'autres ressources (*référence, prérequis, etc.*). Ces annotations sont gérées et corrigées par des experts en se référant à un thesaurus du domaine des mathématiques, ce qui peut affecter la qualité des annotations introduites. Dans cette distinction, citons aussi le système SYFAX (Smei et Ben Hamadou, 2005) qui propose plusieurs annotations désignant : (1) le type de documents (*TD, TP, etc.*) à partir de l'ontologie « Type de documents » et (2) les concepts du domaine traités par le document en se référant à une ontologie du domaine de l'informatique qui nécessite un enrichissement. Notre système ADoP se place dans cette perspective.

Nous remarquons que l'annotation, proposée par les systèmes présentés ci-dessus, est soit manuelle, soit semi-automatique en faisant référence à des ontologies qui nécessitent des ressources énormes pour les gérer. Nous nous sommes intéressés plutôt à l'annotation à la fois automatique et sémantique des documents et nous proposons ainsi une méthode d'annotation de documents pédagogiques basée sur une analyse sémantique des éléments discursifs du texte. L'originalité de notre approche d'analyse des documents pédagogiques revient à se donner les moyens d'accéder au contenu sémantique des textes, pour mieux repérer des séquences particulièrement pertinentes par rapport à un point de vue de l'utilisateur. Par la notion de point de vue, nous désignons le soulignement de certains segments textuels qui focalisent l'attention du lecteur en lisant un document pédagogique (*Définition, Exemple, Exercice, etc.*). Ce soulignement est reproduit par l'exploration contextuelle d'un document pédagogique afin d'expliciter les segments textuels reflétant à un point de vue particulier. Ceci fait l'objet de la section suivante (cf. Section 3).

3 Des points de vue d'annotation de documents pédagogiques par Exploration Contextuelle

En consultant un ensemble de documents pédagogiques, un utilisateur peut être intéressé, par exemple, par « des exemples de requêtes SQL », « des exercices sur le langage UML », « d'un cours sur les réseaux », etc. Ces besoins sont formalisés sous forme de points de vue

(*Définition, Exemple, Exercice, Plan, Cours, Caractéristique, Auteur, Objectif*) qui visent une lecture focalisée et chacun d'eux est explicitement indiqué par des marqueurs linguistiques identifiables dans les textes et caractéristiques d'une intention pragmatique de l'auteur d'un document pédagogique. Cette intention peut être traduite en un point de vue qui doit être identifié par le système pour annoter sémantiquement et discursivement des segments textuels. C'est le principe de la technique d'exploration contextuelle des textes (Desclés, 97) que nous avons employée qui fait appel à des marqueurs discursifs (morphèmes, mot, expression, etc.) explicitant chaque point de vue. Ces marqueurs sont de deux types : des indicateurs et des indices. L'identification des indicateurs est nécessaire au déclenchement d'un ensemble de règles relatives à un point de vue choisi par l'utilisateur. Ces règles doivent identifier dans le contexte Gauche et/ou Droit des indices linguistiques qui orientent vers la confirmation ou l'infirmation de la prise de décision. Ce qui conduit respectivement à attribuer ou non des annotations sémantiques relevant du point de vue choisi. De tels marqueurs linguistiques restent indépendants du domaine : Ils fonctionnent aussi bien en informatique qu'en marketing. Le tableau suivant (cf. TAB.1) présente quelques exemples de points de vue déterminés à partir de notre corpus composé d'un millier de documents de nature principalement pédagogique, une désignation de chaque point de vue ainsi que des exemples illustratifs de structures langagières relatives à chaque point de vue.

Point de vue	Désignation	Exemples de structures langagières
Définition	Une définition d'une notion dans un contexte pédagogique	«...est défini par…», «…se définit comme….», «Par définition, …»
Exemple	Les exemples énoncés dans un document pédagogique	«Par exemple, …», «Un exemple….est illustré……», «Nous donnons un exemple de …»
Exercice	Les exercices de tous types (TD, TP, etc.)	«Exercices sur…», «Ecrivez…», «Travaux dirigés…»

TAB. 1 - *Des exemples de points de vue pour annoter des documents pédagogiques*

Chaque point de vue est décrit de la manière suivante : (1) Un ensemble de classes et de sous classes d'indicateurs et d'indices, (2) un ensemble de règles où chaque règle relie une classe d'indicateurs à différentes classes d'indices. Ces règles ont été développées dans le cadre de notre travail. Des exemples de règles relatives aux points de vue «Définition», «Exemple», «Exercice» sont présentées dans le tableau suivant (cf. TAB. 2).

Règle		R-Définition3	R-Exemple4	R-Exercice2
Point de vue		Définition	Exemple	Exercice
Indicateur	Valeur	défini.txt	exemple.txt	question.txt
	Contexte	Phrase	Phrase	Phrase
Premier indice	Valeur	par-comme.txt	voici.txt	-----
	Espace	Droite	Gauche	------
Annotation	Valeur	«Définition»	«Exemple»	« Exercice »
	Fiabilité	Forte	Forte	Moyen
	Espace	Phrase	Phrase	Phrase

TAB. 2 - *Des exemples de règles relatives aux points de vue « Définition », « Exemple », « Exercice »*

4 Méthode proposée pour l'annotation des documents pédagogiques

Un premier travail préalable de traitement linguistique a été effectué à partir d'un corpus initial pour constituer les ressources linguistiques. Il s'agit (1) d'établir un ensemble de listes de marqueurs linguistiques qui expriment chaque point de vue, et (2) de constituer les règles permettant d'identifier les segments textuels qui relèvent de ce point de vue. Une fois ce travail effectué, notre méthode d'annotation sémantique de documents pédagogiques comporte une étape préalable de prétraitement qui consiste en une segmentation de tous les documents du corpus en unités linguistiques qui sont supérieures ou inférieures à la phrase normative. Pour ce faire, nous avons utilisé l'outil de segmentation SegaTex réalisé par Mourad (2001) pour son aspect multilingue et sa disponibilité (cf. FIG. 1).

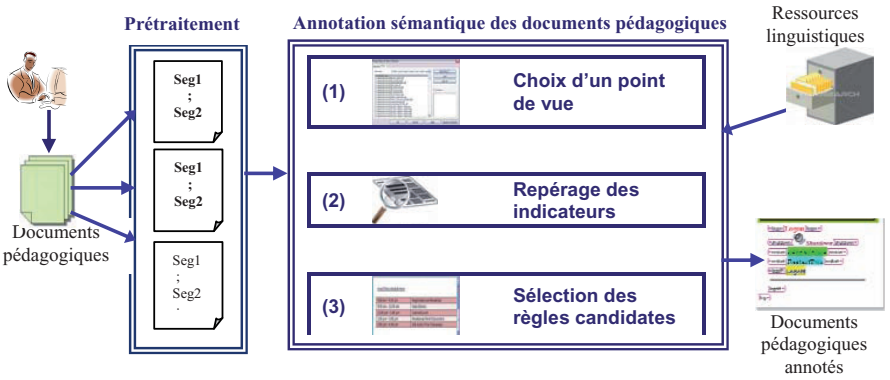

FIG. 1 - *Méthode proposée pour l'annotation de documents pédagogiques*

4.1 Choix d'un point de vue

Lors du lancement du processus d'annotation par notre système ADoP, l'utilisateur doit choisir un point de vue. Cette étape permet de sélectionner, à partir des ressources linguistiques, l'ensemble des règles relatives au point de vue choisi. Par exemple, si l'utilisateur choisit le point de vue «Définition», seules les règles régissant ce point de vue sont appliquées dans le processus d'annotation. Ainsi, le résultat produit l'ensemble des règles relatives au point de vue choisi par l'utilisateur.

4.2 Repérage des indicateurs

Après la sélection de l'ensemble des règles relatives au point de vue choisi, nous avons donné la possibilité à l'utilisateur de choisir une règle à appliquer sur le corpus. Une fois ce choix effectué, le système passe à un repérage des indicateurs relatifs à la règle choisie. Il s'agit de lancer une recherche de chacun de ces indicateurs dans les différents segments

constituant chaque document pédagogique. Si le système identifie au moins un indicateur dans l'espace de recherche, il passe à l'étape suivante du processus d'annotation qui consiste à la sélection des règles candidates (cf. Section 4.3). Dans le cas où aucun indicateur de la présente règle n'a été identifié, le système passe à l'application de la règle suivante qui sera choisie par l'utilisateur et qui est relative au même point de vue.

4.3 Sélection des règles candidates

L'identification d'un indicateur permet de retenir la règle courante comme une règle « Candidate ». Le système passe alors à la vérification de la présence ou de l'absence des indices relatifs à cette règle. En cas où les indices sont vérifiés, la règle en cours devient « Applicable » pour annoter le segment analysé. Sinon, la procédure d'annotation par la règle en cours est annulée. Pour chaque annotation appliquée, le segment annoté est stocké dans une base de données sous forme d'un enregistrement composé des champs suivants : l'emplacement du document qui contient le segment en question, le segment annoté, le point de vue annotant le segment (*Définition, Exemple, Exercice, cours, etc.*) et le titre du paragraphe annoté.

5 Expérimentation et Résultats

Pour valider notre approche, nous avons développé le système ADoP. Il comporte le module de segmentation des textes en utilisant l'outil de segmentation SegaTex, et le module annotation des documents selon différents points de vue. Nous avons utilisé un corpus constitué de 1000 documents, principalement de nature pédagogique (*support de cours, Travaux dirigés, Travaux pratiques, etc.*) et nous avons appliqué l'ensemble des règles d'exploration contextuelle sur ces documents. Quelques résultats d'annotation sont présentés dans le tableau suivant (cf. TAB. 4).

Point de vue	Nbre de segments annotés	Nbre de segments annotés correctement	Précision	Rappel
Définition	228	140	61.4%	93%
Exemple	357	349	97.8%	95%
Exercice	760	705	92.8%	98%

TAB. 4 - *Résultats de l'expérimentation de l'annotation avec ADoP*

6 Conclusion

Nous avons développé un système d'annotation de documents pédagogiques ADoP qui se base sur une approche d'analyse sémantique des éléments discursifs du texte pour annoter automatiquement un document. Les expérimentations sont réalisées sur un corpus composé principalement de documents pédagogiques et les résultats obtenus sont prometteurs. Dans la continuité de ce travail, nous proposons l'enrichissement des ressources pédagogiques par d'autres points de vue d'annotation (*méthode, Date de création, etc.*) et l'extension de notre système à un système de recherche d'informations à partir de documents pédagogiques.

Références

Bodain, Y. (2006). Logiciel d'annotation pour la conception de cours sur le web sémantique. Montréal : IHM.

Buffa, M., S. Dehors, C. Faron-Zucker, and P. Sander (2005). Vers une approche Web Sémantique dans la conception d'un système d'apprentissage. Revue du projet TRIAL SOLUTION , WebLearn,.

Dehors, S., C. Faron-Zucker, J.P. Stromboni and A. Giboin (2005). Des annotations Sémantiques pour apprendre : l'Expérimentation QBLS. WebLearn.

Desclés, J.P. (1997). Système d'exploration Contextuelle, Co-texte et calcul du sens, 215-232.

Desclés, J.P. et B. Djioua (2007). La recherche d'informations par accès aux contenus sémantiques : vers une nouvelle classe de systèmes de recherche d'informations et de moteurs de recherche (Aspects linguistiques et stratégiques). Revue Roumaine de Linguistique, Tome LII, N° 1-2.

Elkhlifi, A. et R. Faiz (2009). Automatic Annotation Approach of Events in News Articles. International Journal of Computing & Information Sciences, Vol. 6, N°1,19-28.

Faiz, R. (2006). Identifying relevant sentences in new articles for event information extraction. International Journal of Computer Processing of Oriental Languages, (IJCPOL), World Scientific, Vol. 19, No. 1, pp. 1-19, 2006.

Kiryakov, A., B. Popov, I. Terziev, D. Manov et D. Ognyanoff (2004). Semantic Annotation, Indexing and Retrieval. Journal on web semantics.

Mourad, G. (2001) Analyse informatique de signes typographiques pour la segmentation de textes et l'extraction automatique des citations. Réalisation des applications informatiques : Segatex et CitaRE. Thèse de Doctorat, Université Paris-Sorbonne.

Smei, H.et A. Ben Hamadou (2005). Un système à base de métadonnées pour la création d'un cache communautaire-Cas de la communauté pédagogique. IEBC : Tunisie.

Teissedre, Ch. (2007) La définition, Etude linguistique, Utilisation dans un système de Recherche d'Information, Comparaison avec le « Define » de Google. Master1-Informatique et Ingénierie de la langue pour la gestion de l'Information. Université de la Sorbonne, Paris IV, France.

Crozat, S. et P. TRIGANO (2002). Structuration et scénarisation de documents pédagogiques numériques dans une logique de massification. Revue STE (Sciences et Techniques Educatives), vol.9, N°3, Ed° Hermès.

Summary. The use of pedagogical documents available on the web becomes increasingly broad both for the teacher who needs to prepare his support of course than for the student who wishes to, for example, autoform. The description of a pedagogical document feeding by metadata is a solution that adds value to the document to clarify the information placed in this document. In this context, we propose a method of annotation of pedagogical documents according to various points of view, which is based on the semantic analysis of the discursive elements of the text.